Research Data Management in India: A Pilot Study

Nishtha Anilkumar[1,⋆]

[1] *Physical Research Laboratory, Ahmedabad, India*

Abstract. Data is a by-product of the research process where published results are the output. More and more research institutes are getting interested in this by-product. Data could be in the form of statistics, experimental results, observational data, interview recordings, etc. Organizing the varied forms of data is a challenge for any institute. This is particularly so in the present scenario of constantly changing technologies for data storage and retrieval. The funding agencies are making it mandatory to archive these datasets so that these can be preserved for the posterity and / or re-used by others. Libraries, as a part of the research institutes, seem to be well equipped to organize and manage these datasets. The author undertook the present study to find the level of involvement of libraries in 'data management' in India. A survey was done to assess the awareness about data curation, data archival policies, infrastructure required, technologies used, etc. The survey sample consisted of 15 national research / academic institutes in India. The study showed that libraries' role in data management in research / academic institutes was still at a very early stage of development in India.

1 Introduction

Research data management (RDM), also sometimes known as research data curation, essentially involves collecting and organising data which is part of the research outcome in such a manner so as to facilitate easy access and re-use. Research data covers a broad range of types of information like documents, spreadsheets, field notebooks, diaries, audio tapes, video tapes, images, spectra, models, algorithms, scripts, protocols, workflows, software, standard operating procedures, methods, etc. [1]. Research data can be defined as "the recorded factual material commonly accepted in the scientific community as necessary to validate research findings" [2]. It does not include preliminary analysis, drafts of scientific papers, and plans of future research, peer reviews, communication with colleagues, trade secrets, commercial information, personnel and medical information. Research data management also includes record management of items like project files, grant applications, ethics applications, technical reports, research reports, signed consent forms, correspondence regarding these matters.

2 Earlier studies

As a result of data gaining importance, academic and research institutes are developing infrastructures and services to support the management of research data on their campuses [3]. In 2012 data

⋆e-mail: nishtha@prl.res.in ORCID 0000-0002-6091-2074

curation was identified as one of the top trends in academic libraries [4]. A study carried out by Akers, et al [5] surveyed eight universities in the United States about how they developed programs for supporting RDM. The study concentrated on the role of the library in educating and assisting the researchers in managing the data before, during, and after their research projects are done. It identified similarities and differences in initial motivation to provide RDM, collaborative relationships with other sections/divisions, approaches to assess the users' needs and changes in competencies required in library staff to carry out RDM. Another study carried out by Norman & Stanton [6] illustrates an evolution toward a strategic vision for library leadership in supporting RDM at the University of Sydney by exploring three stories. Each story throws light on key ingredients that characterize RDM support – researcher engagement, partnerships with other divisions and the complementary roles of policy and practice.

3 Need to preserve/manage research data

Research data is an important and expensive output of the scholarly research process across all disciplines. It is an essential part of the evidence necessary to evaluate research results, and to reconstruct the events and processes leading to them. Its value increases as it gets aggregated into collections and it becomes more available for re-use to address new and challenging research questions. Without proper organization, this value is greatly diminished.

With increasing numbers of 'born digital' research publications, there are new possibilities to store and preserve data with the benefits of making research more usable and replicable, saving the resources, and fostering inter-disciplinary research. The future lies in data being easily and effectively stored, preserved, shared, discovered and re-used in support of researchers and for advancement of science.

Most data are produced or gathered as part of publicly funded research, so it needs to be transparent, accountable and available. Currently in many countries, major funding agencies mandate that applicants submit a data management plan (DMP) as part of their research proposal. The data life cycle provides an overview of the stages involved in successful management and preservation of data for use and reuse. The following image of the data cycle clearly shows the importance of data curation for preserving and re-using and taking the research forward.

4 Stakeholders in RDM

RDM consists of a number of different activities and processes associated with the data lifecycle involving the design and creation of data, its storage, security, preservation, retrieval, sharing and reuse. A number of professional services, including academic support, library services and computing services, have an important role to play in RDM at the institutional level. Erway [7] has identified three main stakeholders in carrying out RDM in any institute. These are Researchers, Library & Information Services and IT Services.

4.1 Researchers

Researchers as the producers of the research data are the first stakeholders. The researchers are usually faced with a mix of requirements – managing their data and open access mandates from their funding agency or parent organization. This makes them reluctant to engage in data sharing. Any endeavour to share that data will depend on the trust they have with the RDM unit. Researchers should be explicitly informed of resulting decisions and procedures so that the trust level rises and they become more open to sharing their data for preservation and access.

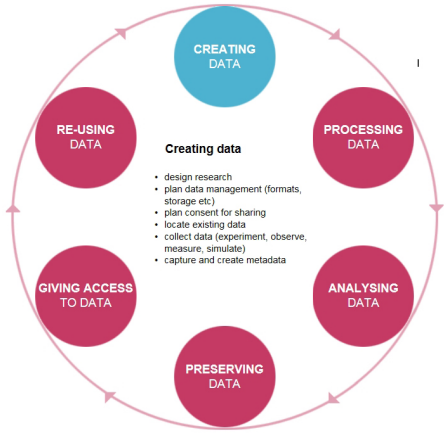

Figure 1. Data life cycle according to UK Data Archives - Posted by: Matthias Töwe, 21/04/2016

4.2 IT Services

In present times most research institutes have robust IT infrastructure which could support advanced data acquisition, storage, management, security, integration, mining and visualization as well as other information processing services. Currently, most researchers keep irreplaceable data on personal storage devices without documentation, version control or back up. Even where data is preserved effectively, it is not being made available to others for re-use. All infrastructure must include systems for documenting, depositing, managing, archiving and preserving data facilitating efficient search and retrieval and providing access.

4.3 Library & Information Services

Libraries are very well positioned to carry out RDM because they understand the need for standards-based information organization and because of the information management skills that librarians have, in particular assigning metadata to the information item for easy retrieval. Besides, libraries have wide experience in collection building, creating and maintaining institutional repositories, preservation and access to library material, including print as well as digital. There is also a potential connection between RDM and the open access agenda that libraries are actively promoting. In addition, libraries have extensive networks with other academic / research divisions within the institute and also with libraries of other institutes. Libraries have earned a trusting relationship with researchers, as their publications are now being preserved in the institutional repository by the library. The general perception of the library is as a safe, sustained and trusted unit for long term document/data preservation. The infrastructure used for an IR may be scaled up to accommodate the data sets. Also, most libraries have experience with copyright issues related to ownership of both source materials and research publications. Libraries are also best suited for the dissemination of information. Libraries can also provide help to researchers with depositing their data to international subject repositories.

An important component of RDM is advice and training of users apart from the technical infrastructure. This helps the librarians to actively re-invent their role and develop while supporting RDM.

Through these activities libraries can play a key role in carrying out RDM effectively. Auckland [8] has identified the following activities which libraries can undertake for data management:

- offering advice / information on funding sources
- conducting literature reviews or CA alerts for research projects
- bibliometric and impact measurement
- reference management software training like Mendeley, Zotero, etc.
- advocacy for open access / institutional repository
- data analysis advice
- advice on copyright issues
- advice on archiving of research records such as correspondence

5 Challenges in RDM

The use of powerful computing technology across disciplines now means that a greater number of researchers generate and use large datasets as part of the research process [9]. Storing this data in a form that can be easily accessed, processed and analysed is a very challenging activity for any research/academic institute. Also, the datasets are potentially fragile, being vulnerable to storage failures and technological obsolescence. It may also be sensitive containing personnel or health information, and so it needs to be managed with proper security measures. The other activities, like reformatting for analysis in various software packages, processing them for potential re-use and carrying out various preservation activities on them, also pose challenges. The responsibility for addressing these challenges then falls on higher education institutes or research institutes. Existing technical infrastructure can be coordinated to support data management. Robust infrastructure needs to be developed to support researchers to manage their data effectively.

One of the major challenges for librarians in taking up RDM for researchers is capacity and workload on the existing, and shrinking, staff. Other key issues to consider are the cost of RDM, including infrastructure, persuading the scientists to deposit their data for preservation and sharing, and convincing the management that their information handling skills are relevant for data curation.

6 Data Policy

Making data available for preservation and re-use adds value to published articles. This preserved data may be used for entirely new purposes in the future, thus increasing the Return on Investment (ROI) of the funding allocated to a particular research project. It could lead to collaborations and new areas of research. Hence, most of the funding agencies have now mandated submission of data management plans along with the research proposal submissions by researchers. Many universities which have no funder mandates for DMPs still carry out data curation and have a data policy in place simply because it is a good practice. Having a data policy helps in clarifying many issues and acts as a guide for the staff carrying out RDM. Before framing a data policy, an institute should constitute a committee comprising of various stakeholders and chaired by a director/dean. Briney, et al [10] have iterated that following points need to be explicitly mentioned while framing a data policy:

1. Why data repository is important?
2. who owns the data?

3. Who is responsible preserving and organising the research data?
4. In case of collaborative research, who will preserve the data?
5. what data should be retained? Who decides which data to keep? No institute can or should retain all research data generated by its researchers. As data curation requires extensive staff time and financial resources, it should be ensured that these are spent on data worth keeping only.
6. How open should the data be? Will it be available immediately or will there be an embargo of 2-3 years and then made open?
7. How long the data is to be preserved? Data may have long-term scientific or institutional value, but all preserved data should be reviewed. When an agreed upon retention period is due to expire, how will it be decided, whether it should be extended?
8. When a dataset is considered no longer to be worth keeping in the repository, who should be notified? Should these weeded out datasets be offered to others or destroyed? What records should be kept to document the weeding out of the dataset?
9. How should digital data be preserved? Each dataset will have unique preservation need.
10. What steps will be taken to preserve the data so that hardware obsolescence does not lead to data loss?
11. Should the DMP be kept with data? What other information should be provided, such as project and personnel records or documentation of instrument calibration?
12. Which kind of metadata is required for different types of data sets?
13. Which software will be used for retrieving the stored data sets?
14. Are the file formats of the data supported by the repository? What descriptors should be applied?
15. What are the implications of cloud storage for data preservation?
16. What are the ethical issues?
    - how will sensitive data be identified and contained?
    - have the consent forms been designed for allowing to share the data after the embargo period say 2-3 years?
17. How is the data accessed?
    - is it necessary to only make the metadata discoverable, with links to the data files?
    - how will the repository monitor access to ensure that restrictions are being enforced?
18. How will the costs be managed? RDM will incur substantial costs.
    - where will the funding come from? if the funding is project based, then how will long term preservation be supported?
19. What happens when the primary researcher leaves the institution?
20. How will the data be acknowledged and cited?

Table 1. Data archiving at various institutes

Institute	Data archiving	Section responsible	Data policy
Bose Institute, Kolkata	Yes	Library	being planned
INFLIBNET, Gandhinagar	Yes	other	being planned
IIAP, Bangalore	Yes	other	no policy
IIM, Ahmedabad	No	Library (being planned)	no policy
IISER, Pune	No	No	no policy
IIT, Gandhinagar	No	No	no policy
IPR, Gandhinagar	No	Library (being planned)	no policy
IUCAA, Pune	No	No	no policy
NIO, Goa	Yes	other	Yes
NRSC, Hyderabad	No	No	no policy
RRI, Bangalore	Yes	Library	no policy
SAC, Ahmedabad	Yes	other	no policy
SINP, Kolkata	No	No	no policy
TIFR, Mumbai	No	No	no policy
Nirma University, Ahmedabad	No	No	no policy

7 Present study

The Physical Research Laboratory, Ahmedabad, to which I am affiliated, wished to start a data curation project for posterity. To assess the level of data curation in India, I decided to do a survey of mainly research institutes and a few institutes of higher education. I sent a questionnaire to 25 research institutes out of the 166 research institutes in India. These 25 institutes were representative members of FORSA, DST, DAE and ISRO Consortium. Out of 25 research institutes contacted 11 have responded. These are Bose Institute, Kolkata, INFLIBNET, Gandhinagar, IIAP (Indian Institute of Astrophysics), IPR (Institute of Plasma Research), IUCAA (Inter-University Centre for Astronomy & Astrophysics), NIO (National Institute of Oceanography), NRSC (National Remote Sensing Centre), RRI (Raman Research Institute), SAC (Space Application Centre), SINP (Saha Institute of Nuclear Physics), TIFR (Tata Institute of Fundamental Research). In addition, I also sent the questionnaire to five academic institutes and out of these four responded - IIMA (Indian Institute of Management, Ahmedabad), IITGN (Indian Institute of Technology, Gandhinagar), IISER, Pune and Nirma University, Ahmedabad. The data compiled from the questionnaire provided information about whether data was being archived at all in the institute. If it was, information was given about which section was responsible for archiving it and whether a data policy was formed to carry out data curation. The following table throws light on these indicators:

Out of the fifteen responses received, nine librarians responded that data is not being archived by the library, but data is maintained by individual researchers. Out of these nine institutes, two libraries plan to start this service for researchers very soon (IIMA & IPR).

In four institutes, data is archived by other divisions like Data & Information Division. Looking at these 4 institutes, Inflibnet, IIAP, NIO and SAC, only NIO has a data policy in place.

In two institutes, data is archived in the library but at a nascent stage. A data policy is being formed at present (Bose Institute & RRI).

The above survey results show that in India RDM carried out by libraries is almost non-existent or at the very least on the idea plane only. As part of the study, I also tried to find standalone data centers funded by the Government of India. I found quite a few data centers in India dedicated to

different subject disciplines and catering to researchers, the funding agencies and the general public. These are:

1. ESSO-INCOIS, Hyderabad http://www.incois.gov.in/
   ESSO-INCOIS was established as an autonomous body in 1999 under the Ministry of Earth Sciences (MoES) and is a unit of the Earth System Science Organization (ESSO). ESSO- INCOIS is mandated to provide the best possible ocean information and advisory services to society, industry, government agencies and the scientific community through sustained ocean observations and constant improvements through systematic and focused research. ESSO-INCOIS has been designated as the National Oceanographic Data Centre by IOC/IODE of UNESCO and is also identified as the Regional Argo Data Centre for the Indian Ocean.]ESSO-INCOIS was established as an autonomous body in 1999 under the Ministry of Earth Sciences (MoES) and is a unit of the Earth System Science Organization (ESSO). ESSO- INCOIS is mandated to provide the best possible ocean information and advisory services to society, industry, government agencies and the scientific community through sustained ocean observations and constant improvements through systematic and focused research. ESSO-INCOIS has been designated as the National Oceanographic Data Centre by IOC/IODE of UNESCO and is also identified as the Regional Argo Data Centre for the Indian Ocean. Its main activities are:

   - Provides round-the-clock monitoring and warning services for the coastal population on tsunamis, storm surges, high waves, etc. through the in-house Indian Tsunami Early Warning Centre (ITWEC). The Intergovernmental Oceanographic Commission (IOC) of UNESCO designated ITWEC as a Regional Tsunami Service Provider (RTSP) to provide tsunami warnings to countries on the Indian Ocean Rim.
   - Short term (3-7 days) Ocean State Forecasts (waves, currents, sea surface temperature, etc.) are issued daily to fisher folk, the shipping industry, the oil and natural gas industry, the Navy, the Coast Guard, etc. These forecasts inform users about the expected sea conditions during the next few days and help them to plan their activities at sea.
   - Deploys and maintains a suite of Ocean Observing Systems in the Indian Ocean to collect data on various oceanic parameters to understand the processes in the ocean and to predict their changes.
   - Conducts systematic quality checks and archives all observational, satellite and other oceanic data at the ESSO-INCOIS Data Centre and then makes such data available to students, researchers and any other users.

2. ICRISAT Dataverse Network http://dataverse.icrisat.org/
   ICRISAT performs crop improvement research, using conventional as well as methods derived from biotechnology, on the following crops: Chickpea, Pigeon pea, Groundnut, Pearl millet, Sorghum and Small millets. ICRISAT's data repository collects, preserves and facilitates access to the datasets produced by ICRISAT researchers to all users who are interested in. Data includes Phenotypic, Genotypic, Social Science, and Spatial data, Soil and Weather.

3. ICSSR Data Service (INFLIBNET) http://www.icssrdataservice.in/
   The "ICSSR Data Service" is culmination of signing of Memorandum of Understanding (MoU) between Indian Council of Social Science Research (ICSSR) and Ministry of Statistics and Programme Implementation (MoSPI). The MoU provides for setting-up of "ICSSR Data Service: Social Science Data Repository" and host NSS and ASI datasets generated by MoSPI. The ICSSR Data Service is set-up with an aim to support researchers, teachers and policymakers who heavily rely on high-quality social and economic data for their research. The ICSSR Data

Service extracts and transforms the data from raw datasets before uploading it in the data repository with necessary documentation for the benefit of researcher. Users can explore collection of datasets accompanied with user guides and supporting materials using search interface of the ICSSR Data Service. All datasets documentation and resources are freely available through this platform. The project on setting-up of ICSSR Data Service is being executed by the INFLIBNET Centre with funding from ICSSR.

4. India Meteorological Department http://www.imd.gov.in/
India Weather Portal has an objective to take meteorological observations and provide current and forecast meteorological information for optimum operation of weather-sensitive activities like agriculture, irrigation, shipping, aviation, offshore oil explorations, etc. Users can check weather of various Indian states, national weather, and weather forecast for agriculture, aviation, and ocean services, etc. It gives warning against severe weather phenomena like tropical cyclones, dust storms, heavy rains and snow, cold and heat waves, etc. which cause destruction of life and property.

5. Indian Oceanographic Data Centre (NIO) http://www.nio.org/iodc/
The Indian Oceanographic Data Centre (IODC) was established in 1964 in NIO, Goa. The IODC assists national and international users in developing and enlarging their competence in the field of marine science. IODC plays a dual role, one by dissemination of data / information to the user communities and the other is to assist the data personnel in processing, validating, reformatting different types of data generated from the Indian Ocean region. The main responsibilities of IODC are:

   - acquire oceanographic data and information for the Indian Ocean
   - reformat, and perform quality control checks on data
   - develop /update databases
   - develop value -added data / information products
   - disseminate data / information to users' community, and
   - provide training in oceanographic data/ information management

6. Indian Space Science Data Center (ISSDC) https://www.issdc.gov.in/
Indian Space Science Data Center (ISSDC) is the primary data centre for the payload data archives of Indian Space Science Missions. This data center, located at the IDSN campus in Bangalore, is responsible for the Ingest, Archive, and Dissemination of the payload data and related ancillary data for Space Science missions like Chandrayaan, Astrosat, etc. The Latest news related to the space missions and other activities are also available. The primary users of this facility will be the principal investigators of the science payloads. In addition to them, the data will be made accessible to scientists from other institutions and also to the general public.

7. Kodaikanal Solar Observatory (IIAP) https://www.iiap.res.in/Kodaikanal_Solar_Observatory
The Kodaikanal Observatory of the Indian Institute of Astrophysics is located in the beautiful Palani range of hills in Southern India. It was established in 1899. Solar observations at this observatory over the last 100+ years provide one of the longest continuous series of solar data. Apart from that, simultaneous observations in different wavelengths make this data a unique one and suitable for multi-wavelength studies. The Kodaikanal observatory has been obtaining solar images since 1904 in broad band white light, narrow band Ca-K 393.37 nm and Ha 656.3nm wavelengths. Many of these observations are still continuing. The historical data which were on photographic plates has been digitized. The first level calibration of the Ca-K, white light

and Ha images have been completed and the data is now available through this portal. Final calibrated images and the data will be posted on this portal in near future. The digitised data are available for use by the scientific community.

8. National Informatics Centre http://www.nic.in/
NIC, under the Department of Information Technology of the Government of India, is a premier Science and Technology Organization, at the forefront of the active promotion and implementation of Information and Communication Technology (ICT) solutions in the government. NIC has spearheaded the e-Governance drive in the country for the last three decades building a strong foundation for better and more transparent governance and assisting the government's endeavour to reach the unreached. With the increased expectations from citizens for online services and the number of e-Governance Projects being launched by the Government, the Data Centre requirements are growing exponentially. There is a need to set up strategic infrastructure that facilitates high availability, quick scalability, efficient management and optimized utilization of resources. To fulfil this requirement, NIC has set up state-of-the-art National Data Centres at NIC Hqs, Delhi, Pune and Hyderabad and 30 small data centres at various state capitals to provide services to the Government at all levels. These Data Centres combine round-the-clock operations and management of systems with onsite skilled personnel.

9. World Data Centre for Geomagnetism, Mumbai http://wdciig.res.in/
This data centre has functioned as a division of the Indian Institute of Geomagnetism, Navi Mumbai since its activities commenced in 1991 in coordination with the International Council of Scientific Unions (ICSU) Panel on World Data Centres. It is responsible for the compilation of final hourly absolute values from nine of the Indian magnetic observatories and deposition of this data to the World Data Centre. In recent years, the centre has prioritized its activities related to digital preservation to ensure digital archiving of magnetic data from the traditional media and also digital conservation of very old hand written/printed data volumes and magnetograms. In view of the scientific importance of data from the Colaba-Alibag Magnetic Observatory, old magnetograms and data volumes are being converted to digital images for long term preservation. In the digital preservation process, the creation of metadata has become an important component in storing information related to old and current scientific records for future use. The centre also hosts a database driven website to make datasets available online to the global scientific community [11].

8 Findings

The survey shows that in India, RDM carried out by libraries is still at the nascent stage of development and will take a few years more to become an integral part of RDM activity in research and academic institutes. The survey results also show that although institutional policy is not in place currently, it is in the process of being formed soon. This indicates that the organizational perspective is changing. However, the prospect of dedicated data centers in different subject fields looks to be quite promising.

9 Conclusion

We as librarians need to convince researchers that data preservation and data access are inter-linked. Preservation helps in access, and active use of data is the best reason for continued preservation. As we all know, researchers are under extreme time pressure. So any new solution would need to be

very easy to use and preferably integrated into their existing workflow to have greater chance of being used. By devising an improved information management solution around research activity, some of the problems that researchers themselves regard as most pressing, could be solved. According to Chad & Suzanne [12] creating better records of researchers and research projects would be important building blocks for a successful RDM system. This study helped to crystallize the idea that forming the data policy is the first step to begin data curation for posterity in any research institute. PRL library has presented a proposal for setting up the infrastructure for data preservation. We are in the process of forming a data policy and are engaging with the researchers to discuss the benefits of data archiving for data access. We have a long way to go, but the beginning is made.. . .

References

[1] C.L. Borgman, Journal of the American Society for Information Science and Technology, **63(6)**, 1059–1078, (2012)
[2] OMB Circular 110, Office of Management & Budget, https://www.whitehouse.gov/omb/circulars_a110
[3] C. Tenopir, B. Birch, S. Allard, ACRL White Paper, (2012)
[4] ACRL Research Planning and Review Committee. *College & Research Libraries News, 73*(6), 311–320, (2012)
[5] K.G. Akers, F.C. Sferdean, N.H. Nicholls, J.A. Green, International Journal of Digital Curation, **9(2)**, 171–191, (2014)
[6] B.Norman, K.V. Stanton, International Journal of Digital Curation, **9(1)**, 253–262, (2014)
[7] R. Erway, Research Data Management Policy. Dublin, Ohio: OCLC Research, (2013)
[8] M. Auckland, RLUK Report, (2012), http://www.rluk.ac.uk/wp-content/uploads/2014/02/RLUK-Re-skilling.pdf
[9] S. Pinfield, A.M. Cox, J. Smith, PLoS ONE, **9(12)**, e114734, (2014)
[10] K. Briney, A. Goben, L. Zilinski, Journal of Librarianship and Scholarly Communication, **3(2)**, p.eP1232, (2015)
[11] M. Doiphode, R. Nimje, S. Alex, Data Science Journal, **12**, WDS85-–WDS88, (2013)
[12] K. Chad, & S. Enright, Insights: The UKSG Journal, **27(2)**, 147–153, (2014)